DOCUMENT RESUME

ED 306 257 TM 013 056

AUTHOR Marsh, Herbert W.; And Others
TITLE Goodness of Fit in Confirmatory Factor Analysis: The Effects of Sample Size and Model Complexity.
PUB DATE 22 Feb 89
NOTE 33p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; *Difficulty Level; *Factor Analysis; *Goodness of Fit; *Mathematical Models; Research Problems; *Sample Size
IDENTIFIERS *Confirmatory Factor Analysis; Parametric Analysis; Population Parameters

ABSTRACT

The purpose of the present investigation is to examine the influence of sample size (N) and model complexity on a set of 23 goodness-of-fit (GOF) indices, including those typically used in confirmatory factor analysis. The focus was on two potential problems in assessing GOF: (1) some fit indices are substantially influenced by N so that tests of the same model based on the same variables for a new sample from the same population are not directly comparable unless N is also held constant; and (2) the inclusion of additional parameters may provide an illusory improvement in fit. For data simulated from each of two different population models, values for 17 of the 23 fit indices were at least moderately influenced by N, and many of these indices failed to control sufficiently for the inclusion of superfluous parameters (i.e., parameters that had zero values in the population model). Four of the indices were relatively independent of N and were not significantly affected by the inclusion of superfluous parameters. The four recommended indices are two measures of fit based on the non-centrality parameter proposed by R. P. McDonald, the widely known incremental (relative) index developed by L. R. Tucker and C. Lewis (1973), and a new incremental index--the McDonald-Marsh Index--that is based on one of McDonald's non-centrality indices. Descriptions of the 23 GOF indices used, 10 graphs, and 5 data tables are provided. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306257

Goodness of Fit in Confirmatory Factor Analysis: The Effects of Sample Size and Model Complexity

Herbert W. Marsh

University of Sydney, Australia

Roderick P. McDonald

Macquarie University, Australia

John Balla

University of Sydney, Australia

22 December, 1987

Revised: 22 February, 1989

Running Head: Goodness of Fit

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

HERBERT W. MARSH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM013056

Goodness of Fit in Confirmatory Factor Analysis:

The Effects of Sample Size and Model Complexity

ABSTRACT

The purpose of the present investigation is to examine the influence of sample size (N) and model complexity on a set of 23 goodness-of-fit indices including those typically used in confirmatory factor analysis. For data simulated from each of two different population models, values for 17 of the 23 fit indices were at least moderately influenced by N, and many of these indices failed to control sufficiently for the inclusion of superfluous parameters (i.e., parameters that had zero values in the population model). Four of the indices were relatively independent of N and were not significantly affected by the inclusion of superfluous parameters. The 4 recommended indices are two measures of fit based on the noncentrality parameter proposed by McDonald (in press), the widely known incremental (relative) index developed by Tucker and Lewis (1973), and a new incremental index called the McDonald-Marsh Index (MMI) that is based on one of McDonald's noncentrality indices.

Goodness of Fit in Confirmatory Factor Analysis: The Effects of Sample Size and Model Complexity

The purpose of the present investigation is to examine the influence of sample size (N) and of model complexity on different goodness-of-fit indices used in confirmatory factor analysis (CFA). In CFA responses to p observed variables by N subjects are summarized by a (p x p) sample covariance matrix and it is hypothesized that the corresponding population covariance matrix can be described by K parameters, namely the factor loadings, the factor variances and covariances, and the residual variances. To the extent that the fitted population covariance matrix Σ derived from a set of (in some sense) best-fitting parameters is similar to the observed sample covariance matrix S, the model is supported. The problem of goodness of fit is how to decide whether Σ is sufficiently similar to S to justify the conclusion that a specific model adequately fits a particular set of data. The present focus is how goodness of fit as assessed with a variety of indices varies with N, the number of cases in the data to be fit, and model complexity as measured by K, the number of parameters estimated in a series of nested models.

The classical form of statistical hypothesis testing is generally inappropriate for evaluation of fit in CFA. Cudeck and Browne (1983) noted that since hypothesized models are best regarded as approximations to reality rather than exact statements of truth, any model can be rejected if the sample size is sufficiently large. From this perspective they argued that it is preferable to abandon the statistical hypothesis testing approach. Similarly, Joreskog and Sorbom argued that statistical hypothesis testing is generally inappropriate because "the statistical problem is not one of testing a given hypothesis (which a priori may be considered false) but rather one of fitting the model to data and to decide whether the fit is adequate or not" (p. I.38-39). McDonald (1985, p. 56) also noted that hypothesis testing is inappropriate for selecting a restrictive model since "all common factor hypotheses are false, because all restrictive hypotheses are false, and they will be proven false by the use of a sufficiently large sample size." In actual application only the "saturated" model can be true. Accordingly, a large number of fit indices have been proposed (e.g., Akaike, 1974; Bollen, 1986; Bentler & Bonett, 1980; Bozdogan, 1987; Cudeck & Browne, 1983; Hoelter, 1983; Horn & McArdle, 1980; James, Mulaik & Brett, 1982;

Joreskog & Sorbom, 1981; Marsh, Balla & McDonald, 1988; McArdle, 1986; McDonald, in press; McDonald & Marsh, 1988; Schwartz, 1978; Steiger & Lind, 1980; Tanaka, 1987; Tanaka & Huba, 1986; Tucker & Lewis, 1973) to facilitate the evaluation of fit and the comparison of alternative models.

Desirable Characteristics of Fit Indices.

The focus of the present investigation is on two potential problems in assessing goodness of fit. First, some fit indices are substantially influenced by N so that tests of the same model based on the same variables for a new sample from the same population are not directly comparable unless N is also held constant. Such an effect of N also makes problematic any guidelines of what constitutes an acceptable fit. Thus, some researchers have developed fit indices that are claimed to be relatively independent of N. Second, the inclusion of additional parameters, particularly when based on a posteriori criteria and tested with the same data, may provide an illusory improvement in fit. Thus, some researchers have developed fit indices that are claimed to compensate for capitalization on chance. From these perspectives, an ideal index of fit would be relatively independent of N, provide an accurate measure of goodness of fit for competing models, vary along a well-defined continuum that is easily interpreted, and control appropriately for model complexity.

Many researchers have examined the effect of N on goodness of fit (e.g., Anderson & Gerbing, 1984; Bearden, Sharma & Teel, 1982; Bentler & Bonett, 1980; Bollen, 1986; Boomsma, 1982; Cudeck & Browne, 1983; Gerbing & Anderson, 1985; Hoelter, 1983; Joreskog & Sorbom, 1981; Marsh, Balla & McDonald, 1988; Marsh & McDonald, 1988) and some have proposed fit indices that are claimed to be independent of N. Marsh, Balla, and McDonald used actual and simulated data to demonstrate that nearly all frequently used indices are substantially influenced by N. Of the more than 30 indices that they considered, the Tucker-Lewis index (TLI) was the only frequently used index that was relatively independent of N.[1]

Researchers have also examined the effect of the number of parameters included in the hypothesized model on goodness of fit (e.g., Akaike, 1974; 1981; Anderson & Gerbing, 1984; Bentler & Bonett, 1980; Boomsma, 1982; Bozdogan, 1987; Cudeck & Browne, 1983; Gerbing & Anderson, 1985; James, Mulaik & Brett, 1982; Joreskog & Sorbom, 1981; Schwartz, 1978; Tucker & Lewis, 1973). Many fit indices are monotonically related to model complexity

as measured by the number of parameters estimated in a series of nested models so that for sample data goodness of fit will continue to improve with the addition of more parameters so long as the df is positive. From this perspective the best fitting model will always be the saturated model with df=0. However, for sample data this improved fit due to the inclusion of additional parameters may be due to capitalization on chance. Furthermore the parameter estimates for a saturated model may be uninterpretable and researchers often seek more parsimonious models that are both theoretically defensible and able to describe their data adequately.

Researchers have approached this problem of evaluating fit in relation to model complexity from different perspectives. For example, James, et al., (1982, p. 155) ask "how efficient is the increase in fit going from the null model with many degrees of freedom to another model with just a few degrees of freedom in terms of degrees of freedom lost in estimating more parameters?" Joreskog and Sorbom (1981, p. I. 40) note that when the change in X^2 is close to the difference in df due to the addition of new parameters, then the "improvement in fit is obtained by 'captializing on chance,' and that the added parameters may not have real significance and meaning." Cudeck and Browne (1983; also see Marsh, 1987) proposed the method of cross-validation to determine the ability of a set of parameter estimates to adequately describe data based on new observations from the same population and to determine the extent to which capitalization on chance has occurred. Cudeck and Browne also demonstrated the use of CAK and CSK (see definition in Appendix 1), indices described by Akaike (1974) and by Schwartz (1978) respectively that were rescaled in terms of FF (see Appendix I), for this purpose. Bozdogan (1987) noted that model selection requires researchers to achieve an appropriate balance between problems associated with overfitting and underfitting the data, and that different fit indices vary in the balance of protection that they offer from these conflicting possibilities. Similarly, McDonald (in press) noted the need to strike a balance between badness of fit and model complexity or, equivalently, between goodness of fit and model parsimony. He further noted that this compromise is not an issue of sampling in that even if the true population were known, an appropriate compromise would still be required.

Cudeck and Browne (1983) examined the joint influence of sample size and model complexity on goodness of fit. They considered the CAK and CSK indices that are a function of the number of estimated parameters. These

indices are possibly a useful indication of fit for comparing competing models that vary in the number of parameters used to describe the same data and have recently received much attention (e.g., Bozdogan, 1987). Cudeck and Browne's results, as well as results by Marsh, Balla and McDonald (1988), show empirically that these indices are substantially influenced by N, and McDonald (in press) demonstrated that this relation was inherent in the mathematical form of the indices. The Akaike index penalized the inclusion of additional parameters less severely than the Schwartz index so that it consistently led to the selection of more complex models (see Bozdogan, 1987). This effect of sample size need not invalidate the use of these indices for purposes of model selection if the effects of N are relatively constant across the different models. That is, the same model may be selected as "best" for each of the different sample sizes even though the actual values of the fit indices varied according to sample size. However, Cudeck and Browne found that the relative fit of competing models did vary with N. For small sample sizes, simple models positing fewer parameters had better fit indices whereas for large sample sizes more complicated models positing more parameters, and, ultimately, for sufficiently large sample sizes, the saturated model, had better fit indices. As noted by McDonald (in press), two studies differing only in sample size would on average lead to the support of models differing in complexity and no investigator would reasonably use such indices if the sample size were large enough to require the selection of an uninterpretably complex model.

The Present Investigation

Our objective is to examine the effect of model complexity and of N on a set of 23 goodness of fit indices. Data were generated from one of two known population models and a variety of models used to fit the data were developed in relation to these known population models. Some models posited parameters to be zero that were known to be non-zero for the population model, thus providing models that were under-fit. Other models estimated values for superfluous parameters that were known to be zero for the population model, thus providing models that were over-fit. Covariance matrices to be fit by the alternative models were based on one of six different sample sizes varying from 50 to 1600.

The set of 23 goodness of fit indices considered here are described in more detail in Appendix I. For present purposes the indices are classified

into three types, namely: (a) stand-alone (absolute) indices, (b) type-1 incremental (relative) indices, and (c) type-2 incremental (relative) indices. The 15 stand-alone indices are based on the results of just a target model, the a priori model posited by the researcher to fit the data. These indices are provided by, or easily computed from results provided by, LISREL and most other statistical packages used to fit structural equation models. The incremental indices are based on the difference between the target model and an alternative model such as a "null" model in which Σ is a diagonal matrix (Bentler & Bonett, 1980). Incremental type-2 indices incorporate an expected value of an index for a true model whereas incremental type-1 indices do not (see Appendix). Marsh, Balla and McDonald (1988) examined 19 of the 23 indices considered here -- all but Dk, Mc, Z, and the McDonald-Marsh Index (MMI) -- and found that only the TLI was relatively independent of N (also see footnote 1). McDonald (in press) indicated that his DK and Mc indices were relatively independent of N. Z, because it is monotonically related to X^2, should be affected by sample size. The MMI was developed for purposes of the present investigation.[2]

Method

The CFA Model and Analyses

All analyses were conducted with LISREL V (Joreskog & Sorbom, 1981) using the method of maximum likelihood. In each of the analyses involving 9 observed variables a set of eight substantive models posited between 18 and 33 parameters to define 1, 2, or 3 factors. Hence the df (.5 x 9 x 10 - K) varied from 27 to 12. These eight models and their relation to the population model used to generate the data are summarized in Table 1. A null model was also tested for each covariance matrix such that the reproduced covariance matrix was a diagonal matrix of variances and the nine measured variables were posited to be uncorrelated. The df for the null model (.5 x 9 x 10 - 9 = 36) was constant for all the analyses. These nine models, the eight substantive models and the null model, were tested for each of 120 covariance matrices described below.

Insert Tables 1 & 2 About Here

The Data.

The Sample Sizes. The six sample sizes to be considered in the present

investigation, 50, 100, 200, 400, 800 and 1600, were selected to span the range of sample sizes typically considered in CFA. For each of the two data sets to be considered, ten random samples were generated for each sample size and the same nine models were fit to these 120 (2 data sets x 6 sample sizes x 10 cases) covariance matrices.

Simple structure simulated data (SSIM). The nine measured variables were defined with the random number generator from the commercially available SPSS package (Hull & Nie, 1981). Each variable was defined to reflect only one factor (factor loadings were .6, .7 or .8) and a normally distributed random error component, and the three factors were defined to be correlated (factor covariances were .08, .12, and .24). A total of 31,500 cases were generated and divided into 60 sets of data such that each sample size was represented by 10 covariance matrices. The eight substantive models and the null model were fit to each of the 60 covariance matrices.

The population model used to generate this data was one of the substantive models to be considered (3SF, see Table 1) and thus was the most parsimonious model (i.e., contained the fewest estimated parameters) able to fit the data. Models positing only one or two factors (1UF and 2UF in Table 1) should not be able to fit the data. In each of the remaining five substantive models, all the parameters in the 3SF model are included along with a varying number of additional parameters. These additional parameters are superfluous in that their population values, the values from the population models used to generate the data, are zero. The fit indices of these over-fit models are used to evaluate how various indices are affected by capitalization on chance. To the extent that any of these over-fit models fit the SSIM data significantly better than the 3SF model according to a particular index, then the index does not control for the effects of capitialization on chance. To the extent that any of these models fit the data significantly poorer than the 3SF model according to any particular indices, then, perhaps, the index over-compensates for capitalization on chance. This relation between the substantive models to be tested and the SSIM data was the basis of a priori contrasts used to compare various models (see Table 2).

Complex Structure Simulated Data (CSIM). The nine measured variables were defined as with the SSIM except that six of the nine measured variables -- two for each factor -- were defined such that each should have a small loading (.2) on one factor in addition to the one it was designated to

reflect. (In Table 1, the 9 factor loadings corresponding to those in the SSIM data are called major factor loadings whereas the additional 6 factor loadings in the model used to generate the CSIM data are called minor loadings). Again a total of 31,500 cases were generated and divided into 60 sets of data such that each sample size was represented by 10 covariance matrices, and the null and hypothesized models were fit to these 60 covariance matrices.

The population model used to generate the CSIM data was one of the substantive models to be considered (3CF, see Table 1) and so it is the most parsimonious model able to fit the CSIM data. Model 3UF, positing three unrestricted factors, should also be able to fit the data adequately though it is less parsimonious. Models positing only one or two factors (1Uf and 2UF in Table 1) should not be able to fit the data. Furthermore, in each of the remaining four substantive models positing three factors, either 3 (Models 3F1 and 3F2) or all 6 (Models 3F3 and 3SF) of the minor factor loadings are constrained to be zero. Of these four models, only Model 3F3 contains superfluous parameters, parameters whose population value is zero. This set of models provides additional tests of how the different indices vary according to model complexity and models known to over-fit or under-fit the data in relation to the known population parameters. Two sets of models (Models 3UF and 3CF, and Models 3F3 and 3SF) should be equivalent in their ability to fit the data but differ in the number of parameters that are estimated. For two additional sets of models (3UF vs. 3SF; 3F1 and 3F2 vs. 3SF) the model that should fit best requires more parameters so that an index that over-corrects for capitalization on chance may distort appropriate differences in fit. This relation between the substantive models to be tested and the CSIM data was the basis of a priori contrasts used to compare various models (see Table 2).

Results

The analyses to be described are based on a set of 8 (substantive models) x 6 (sample sizes) ANOVAs which were followed up by the set of 9 a priori contrasts described in Table 2. Separate analyses were conducted for each of the 23 fit indices and separate analyses were conducted for results of the SSIM and CSIM data.

Simple Simulated (SSIM) Data.

Models. In relation to the population model used to generate the SSIM data, Models 3 - 8 should be able to fit the data (i.e., all nonzero population parameters are estimated) whereas Models 1 and 2 should not. For all 23 fit indices there are significant differences in the ability of competing models to fit the data (see Etas attributable to the Model in Table 3), and most of this difference is due to the poorer fits of Models 1 and 2.

For the SSIM data, models 3 - 8 are all able to fit the data but differ in the number of parameters that are posited. Because all these models should be able to fit the data, it could be argued that the models should not differ in goodness of fit. For analysis conducted on just Models 3 - 8 (Table 4) the effect of the model complexity varies substantially with the fit index; 7 indices show significantly better fits when more (superfluous) parameters are estimated, 6 indices show significantly poorer fits when more parameters are estimated, and the remaining 10 indices are not significantly related to the number of estimated parameters.

Insert Tables 3 & 5 and Figure 1 About Here

For all but 3 indices (DK, MC, and LHRI1) the effect of the models interacted significantly with sample size (see Table 3), though the size of this interaction was substantial for only 6 indices. Particularly for these 6 indices there is a similar pattern of interaction. For Models 1 and 2 that are unable to fit the data, fit becomes substantially poorer as sample size increases (see X^2 in figure 1). For Models 3 - 8 that are able to fit the data, differences between models less related to sample size. Thus, for analyses of just Models 3 - 8 (Table 4) the size of this interaction is much smaller. The form of the interaction is illustrated for other selected indices in Figure 1.

Sample Size (N) Effect. The effect of the six levels of N is statistically significant and substantial for 17 of the 23 indices (etas of .26 to .96; see Table 3). For these 17 indices most of this effect can be explained by the linear effect of log N (rs of .25 to .90). The direction of the effect of N, however, depends on the particular index (see Table 3 and Figure 1). The relation between goodness of fit and N is not statistically significant for McDonald's Dk and Mc stand-alone indices and the MMI relative index, and is very small for the TLI relative index.

A Priori Contrasts. The purposes of the a priori contrasts are to test the ability of the 23 indices to differentiate among models known to differ in their ability to fit the data, and to evaluate the indices in relation to capitalization on chance. For the SSIM data, the set of 9 a priori contrasts can be divided into two types. Contrasts 1 and 2 compare models that are known to differ substantially in their ability to fit the data, whereas contrasts 3 - 9 compare models that are all able to fit the data. For contrasts 1 and 2, comparisons based on 20 of the 23 indices are statistically significant and in the right direction. For CN both contrasts are in the right direction but one is not statistically significant. For the two parsimony indices one or both of the contrasts are significant but in the wrong direction. These results based on contrasts 1 and 2 provide support for 20 of the indices, but call into question the usefulness of CN and the two parsimony indices.

Contrasts 3 - 9 are all based on comparisons among Models 3 - 8 that do not differ in their ability to fit the data. Because the SSIM data was generated by a population model containing only 21 parameters estimated in Model 3SF (Table 1), additional parameters are superfluous. For just contrast 5 the models being compared are equally able to fit the data and posit the same number of parameters (each contains 3 superfluous parameters); this contrast fails to reach statistical significance for any of the 23 indices.

Contrasts 3, 4, 6, 7, 8 and 9 all compare models that are able to fit the data but differ in the number of (superfluous) parameters. For each of these contrasts (Table 3), a plus (+) indicates that the model with more parameters fits the data better whereas a minus (-) indicates the opposite. The behavior of the different fit indices in relation to these contrasts vary substantially and fall into three classifications.

1) For 8 indices (FF, LHR, X^2, RMR, GFI, FFI1, LHRI1, X^2I1) all statistically significant contrasts favor the models that posit more parameters. For these 8 indices, even those contrasts that are not statistically significant favor models that posit more parameters. For these indices more complex models positing more parameters fit the data better. Because the true population values for these additional parameters are known to be zero for this simulated data, this improved fit is illusory and due to capitalization on chance.

2) For 6 of the fit indices (CSK, CAK, OCSK, OCAK, PIX^2, and PIRMR), all statistically significant contrasts favor models that posit fewer parameters. For these 6 indices, even those contrasts that are not statistically significant favor models with fewer parameters. That is, models positing more (superfluous) parameters fit the data more poorly than models positing fewer parameters so that these indices can be said to penalize model complexity. The danger in penalizing model complexity too severely is observed for the two parsimony indices in relation to contrasts 1 and 2. For both these contrasts, the better model (in relation to the known population model) posited more parameters. The two parsimony indices so severely penalize the inclusion of additional parameters that better fitting models have significantly poorer indices of fit. Examination of the contrasts for the remaining four indices in this second group suggests that the CSK and OCSK penalize model complexity more severely than CAK and OCAK (also see Bozdogan, 1987, for a mathematical basis for this observation). However, because contrasts 1 and 2 are statistically significant and in the right direction for each of these four indices, there is no basis for claiming that model complexity is penalized too severely. Indeed, it may be reasonable to severely penalize the inclusion of superfluous parameters so long as models better able to fit known population parameters have better indices than models less able to fit known population parameters. Although a useful guideline for simulated data, this condition cannot be tested for real data since the population parameters can never be known.

3) For the remaining 9 fit indices (X^2/df, AGFI, CN, DK, MC, Z, X^2/dfII, TLI, and MMI), none of the contrasts are statistically significant. That is, for these indices models positing more (superfluous) parameters do not differ significantly from models positing fewer parameters.

Summary of SSIM analyses. Analyses of the SSIM data were used to examine the behavior of 23 indices of fit. Four of the indices (DK, Mc, TLI, and MMI) were relatively independent of N and were not significantly affected by the inclusion of superfluous parameters. The remaining 19 indices were at least moderately influenced by N and many were significantly affected by the inclusion of superfluous parameters. CN, in addition to being substantially influenced by sample size, did not differentiate between models known to differ in their ability to fit the data. The two parsimony indices, in addition to being moderately influenced by N, were shown to penalize model complexity too severely.

Complex Simulated (CSIM) Data.

Models. For the CSIM data the effect of the different models is statistically significant and substantial for all 23 fit indices (Table 5). This effect of models interacts significantly with N for 17 of the fit indices, though the size of the interaction is substantial for only 6 indices. The indices most affected by this interaction and the nature of this interaction are similar to that observed for the SSIM data (also see Figure 1), and so are not discussed further.

Insert Table 5

Sample Size (N) Effect. The effect of N is statistically significant and substantial for 19 of the 23 indices (etas of .26 to .96; see Table 5). For these 19 indices most of this effect is linearly related to log N (rs of -.25 to -.90), but the direction of this effect depends on the index (see Table 5 & Figure 1). The relation between goodness of fit and N is not statistically significant for Dk, Mc, TLI and MMI. Again, these results are similar to those observed for the SSIM data.

A Priori Contrasts. For the CSIM data, the set of 9 a priori contrasts can be divided into two types. Contrasts 1, 2, 4, 5, 6, 7, and 9 are between models known to differ in their ability to fit the data, whereas contrasts 3 and 8 compare models that are equally able to fit the data but differ in the number of superfluous parameters that are posited.

Contrasts 1 and 2 are gross tests in that they compare the 3 unrestricted models positing 1, 2 and 3 factors. For 20 of the 23 indices, contrasts 1 and 2 are statistically significant and in the right direction. For CN both contrasts are in the right direction but one is not statistically significant. For the two parsimony indices one or both of the contrasts is significant but in the wrong direction. These results based on contrasts 1 and 2 are similar to findings based on the SSIM data and so are not discussed further.

Contrasts 4 and 9 are also rather gross tests in that models that are able to fit the data (3UF and 3CF) are compared to model 3SF in which all 6 minor factor loadings known to be nonzero in the population model are fixed to be zero. For 17 of the 23 indices, these contrasts are statistically

significant and in the right direction. For CSK, OCSK, OCAK and the two parsimony indices, one or both of these comparisons is statistically significant and in the wrong direction. This demonstrates that with respect to these contrasts, these indices penalize model complexity too severely.

Contrasts 6 and 7 are less gross in that the models being compared differ in terms of only 3 of the 6 minor factor loadings. For 16 of the 23 indices, these contrasts are statistically significant and in the right direction. For the two parsimony indices, both these contrasts are statistically significant but in the wrong direction. For CSK the contrasts are in the wrong direction, but not statistically significant. For CAK, CN, OCSK, and LHRI1, one of these contrasts was not statistically significant though none were in the wrong direction. This demonstrates that with respect to these contrasts, the at least the parsimony indices penalize model complexity too severely.

Contrast 5 compares models 3F1 and 3F2 in which 3 of the 6 minor factor loadings are fixed to be zero with model 3F3 in which all 6 are fixed to be zero. Thus, Models 3F1 and 3F2 should be able to fit the data better than model 3F3. In model 3F3, however, 3 additional superfluous parameters are also estimated so the df is the same for all three models. For only 5 (X^2, X^2/df, OCAK, OCSK, and Z) of the 23 indices is this contrast statistically significant and in the right direction. For all 23 indices, however, this contrast was in the right direction and the contrast approached statistical significance for many of these indices. It is also relevant to note that the results of this contrast are not related to the number of estimated parameters in that all the models posited the same number of parameters.

Contrasts 3 and 8 compare models that are equally able to fit the data but differ in the number of (superfluous) parameters. As observed with the SSIM data in this situation, the behavior of the indices fell into three general categories. For 10 indices one or both of these contrasts are statistically significant such that the model positing more parameters fits the data better. As noted previously, this improved fit is illusory and represents capitalization on chance. For 4 indices one or both of these contrasts are statistically significant such that the model positing fewer parameters fits the data more poorly. For 9 indices, neither of these contrasts is statistically significant.

Summary of CSIM analyses. Analyses of the CSIM data were used to

examine the behavior of 23 indices of fit. Four of the indices (DK, Mc, TLI, and MMI) were relatively independent of N and were not significantly affected by the inclusion of superfluous parameters. The remaining 19 indices were at least moderately influenced by N and many were shown to significantly capitalize on chance when superfluous parameters were estimated. CSK, OCSK, the two parsimony indices, and perhaps CAK in addition to being moderately influenced by N, were shown to penalize model complexity too severely in that models less able to fit the data provided better fits than models better able to fit the data. These findings are generally consistent with those based on the SSIM data.

Discussion

Results for both the SSIM and CSIM data lead to clear conclusions about the behavior of fit indices considered here. For 19 of the indices -- all but Dk, Mc, TLI, and MMI -- there was a moderate or large effect of N. These results are consistent with conclusions by Marsh, Balla and McDonald (1988), Marsh and McDonald (1988), and McDonald (in press). These same 4 indices were also shown to be not significantly affected by the inclusion of superfluous parameters that had population values known to be zero. In contrast, the addition of superfluous parameters resulted in significant improvements in fit that was due to capitalizing on chance for many of the indices. Other indices were shown to penalize model complexity too severely in that inclusion of parameters that had nonzero values in the population led to a significantly poorer fit. In some instances indices penalized model complexity so severely that models better able to fit the data in relation to the known population parameters produced poorer fit indices than models that were less able to fit the data but contained fewer parameters. Whereas a few other indices were not significantly affected by the introduction of superfluous parameters, all of these other indices were at least moderately affected by sample size. Hence, in relation to the desirable characteristics of fit indices considered here, there is clear support for only Dk, Mc, TLI, and MMI indices.

The empirical results presented here suggest little basis for choosing among the four recommended indices. In fact correlations among these indices are .97 or higher for the data considered in this study. Theoretically, however, the four indices differ in important ways. McDonald's two indices are absolute or stand-alone indices that depend only on the model being

tested. Mc may be preferable to Dk in that it varies on a zero-to-one continuum that may prove to be more easily interpreted. McDonald noted, however, that such interpretations must be subjective since only the saturated model is true in application. TLI and MMI are both incremental or relative indices that depend on the fit of a null model as well as the fit of the hypothesized model. The TLI is much better known than the new MMI, but its estimation is frequently unstable particularly when sample size is small (see Figure 1; also see Anderson & Gerbing, 1984; Marsh, Balla & McDonald, 1988). Further research may show, however, that the same problem applies to the MMI although it was not apparent in the present investigation. The Dk, Mc, and MMI also differ from the TLI in that the first three are monotonically related to the number of estimated parameters whereas McDonald and Marsh (1988) show that the TLI can be written as an index of fit that is weighted by a parsimony index. In this respect, the TLI can be said to penalize model complexity whereas the other indices do not. In the present investigation this mathematical distinction between these indices was not demonstrated empirically. This can apparently be explained by the observation that when the TLI is sufficiently large, as in most of the contrasts in the present investigation, the size of this penalty is negligible. Hence, it is possible the these four indices will differ more substantially in other situations and this is an important question for further research.

The present investigation is based on a variety of models fit to simulated data of varying sample sizes derived from only two different population models. Hence, there is concern about the generality of our findings, particularly with respect to use of simulated data. We found that 19 of the indices considered here were at least moderately affected by sample size, and that many of these were significantly influenced by the addition of superfluous parameters that represented capitalization on chance. Other indices were shown to penalize model complexity too severely in that models able to fit the data resulted in significantly poorer indices than models not able to fit the data. These findings call into question the usefulness of these indices as indicators of fit according to the criteria proposed here. Even if other research shows any of these indices to be useful in some specific situations, our results would stand as counterinstances to the generality of claims to their usefulness.

We found that 4 of the indices considered here were relatively

independent of sample size and were not significantly affected by the inclusion of additional parameters. The conclusions about the effect of N on these indices is consistent with other empirical research and mathematical derivations of the indices. Further tests of the generality of our findings based on the data sets considered here, however, will help clarify the relations between these indices and model complexity. With further research it may be possible to establish useful guidelines on the values of these indices that constitute acceptable fit, but such attempts may be unjustified for any of the other 19 indices considered here. For real data, however, none of the population parameters will generally have a zero value so that there may be no rational basis for concluding that any restricted model fits the data better than the saturated model. Ultimately model selection must be based on evaluation of fit, the behavior of competing models, and substantive issues. From this perspective it would be undesirable to establish absolute guidelines about what constitutes an adequate fit that are independent of the research context.

Footnotes

1 -- Several additional incremental (relative) indices referred to as Type-2 incremental indices (see Appendix for discussion of Type-1 and Type-2 incremental indices) by Marsh, Balla and McDonald (1988) were found to be relatively independent of sample size for the four data sets considered in that study. McDonald and Marsh (1988) subsequently showed, however, that by their mathematical form some of these indices should vary with sample size under certain conditions that did not exist in the data sets considered by Marsh, Balla and McDonald (1988).

2 -- McDonald first developed his two indices, Mc and Dk, based on the noncentrality parameter in late 1986, as described by McDonald (in press). Shortly after their development, in February of 1987, Marsh and McDonald proposed the incremental type-1 and type-2 forms of both these indices for purposes of the present investigation. Only the results of the DkI2 are actually presented here. DkI1 and DKI2 are mathematically identical (see Appendix) whereas the pattern of empirical results based on the MCI2 were nearly identical to those based on DKI2. MCI1, because it was significantly related to sample size, was not pursured for purposes of the present investigation. Subsequently, in October 1988, McDonald and Marsh evaluated the mathematical properties of DKI2 more fully in research described in Marsh and McDonald (1988). For purposes of that paper and the present investigation, the index is referred to as the McDonald and Marsh index (MMI).

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

Akaike, H. (1981). Likelihood of a model and information criteria. Journal of Econometrics, 16, 3-14.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Bearden, W. O., Sharma, S., & Teel, J. R. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. Journal of Marketing Research, 19, 425-530.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, Mass: MIT Press.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. Psychometrika, 51, 375-377.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskog and H. Wold (Eds.),Systems under indirect observation: Causality, structure, prediction (Part I). Amsterdam: North-Holland.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 345-370.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18, 147-167.

Gerbing, D. W., & Anderson, J. C. (1985). Effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. Multivariate Behavioral Research, 20, 255-271.

Hoelter, J. W. (1983) The analysis of covariance structures: Goodness-of-fit indices. Sociological Methods & Research, 11, 325-344.

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical/statistical model building (MASMOB) in research on aging. In Poon, L. W. (ed.), Aging in the 1980's: Selected contemporary issues in the psychology of aging. (pp. 503-541) American Psychological Association, Washington, DC.

Hull, C. H., & Nie, N. H. (1981). SPSS update 7-9. New York: McGraw-

Hill.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis. Assumptions, models, and data. Beverly Hills, CA: Sage.

Joreskog, K. G. & Sorbom, D. (1981). LISREL V: Analysis of Linear Structural Relations By the Method of Maximum Likelihood. Chicago: International Educational Services.

Marsh, H. W. (1987). The factorial invariance of responses by males and females to a multidimensional self-concept instrument: Substantive and methodological issues. Multivariate Behavioral Research, 22, 457-480.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 102, 391-410..

McDonald, R. P. (1985). Factor analysis and related methods. Hillsdale, NJ: Erlbaum.

McDonald, R. P, & Marsh, H. W. (1988). Choosing a multivariate model: Noncentrality and goodness-of-fit. (In Review).

McDonald, R. P. (in press) An Index of Goodness-of-fit based on noncentrality. Journal of Classification,.

McArdle, J. J. Latent variable growth within behavior genetic models. Behavior Genetics, 16, 163-200.

Schwartz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Steiger, J. H., & Lind, J. M. (May, 1980). Statistically-based tests for the number of common factors. Paper presented at the Psychometrika Society Meeting, Iowa City.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. British Journal of Mathematical and Statistical Psychology, 46, 621-635.

Tucker, L. R. & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 54, 426-482.

APPENDIX I

Descriptions of the 23 Goodness of Fit Indices Used in This Study

Four types of fit indices are considered here. Stand alone indices are based on results of just the hypothesized model. Two forms of incremental indices, called type-1 and type-2 for present purposes, are based on differences in fit between a hypothesized model and a null model. Parsimony indices are an alternative form of the type-1 incremental indices that impose a penalty function for the inclusion of additional parameters.

I. Absolute, Stand-alone Indices.

The maximum likelihood fitting function (FF) and the scaled likelihood ratio (LHR). Although not typically presented as fit indices (but see Cudeck & Browne, 1983), the FF and LHR are the basis for the X^2 test statistic and most other fit indices. The FF has a minimum value of 0 when E = S, but does not have an upper bound. The scaled LHR has a maximum value of 1.0 when E = S and a minimum value of zero. The FF and LHR are defined as:

(1) $FF = X^2 / (N)$,

(2) $LHR = Exp(X^2/(-2 \times (N))) = e^{-1/2\ FF}$.

X^2 and X^2/df Ratio. These two indices continue to be the most frequently used indices. The X^2 for a false model varies directly with sample size, but the X^2 for a true model does not. In CFA the df does not vary with the sample size, so that the effect of sample size on the X^2/df must necessarily be the same as for the X^2. For alternative models of the same data, increasing the number of parameters necessarily results in a better (i.e., lower) X^2. Because the X^2/df ratio incorporates a penalty function for using more parameters, it may be poorer if additional parameters result in little improvement in X^2. They are defined as:

(3) $X^2 = tr\ (E^{-1}\ S - I) - \log |\ E^{-1}\ S\ | = (N)\ FF$,

(4) $X^2/df = ((N)/df)\ FF$.

LISREL's root mean square residual (RMR). Joreskog and Sorbom (1981, p. I.41) define the RMR as the square root of the mean of squared residuals in S and E. When S and E are based on correlation matrices RMR is strictly bounded by 0 and 1. For covariance matrices RMR still has a lower-bound of zero but does not have an upper bound. Thus RMR must be interpreted in relation to the size of the variances and covariances of the measured variables, and cannot be compared across applications based on different variables. RMR is defined as:

(5) $$RMR = [\ 2 \sum \sum (s_{ij} - e_{ij})^2 / (p \times (p+1))]^{1/2},$$

where s_{ij} and e_{ij} are elements in S and E

LISREL's goodness-of-fit (GFI) and adjusted GFI (AGFI). Joreskog and Sorbom (1981; also see Tanaka & Huba, 1986) describe the GFI and AGFI as computed by LISREL. They state that GFI is "a measure of the relative amount of variances and covariances jointly accounted for by the model" and assert that "unlike X^2, GFI is independent of the sample size" while AGFI "corresponds to using mean squares instead of total sums of squares" (Joreskog & Sorbom, 1981, p. I. 40-41). Thus AGFI incorporates a penalty function for additional parameters. Joreskog and Sorbom suggest that GFI and AGFI will generally fall between 0 and 1, but that it is possible for them to be negative. They are defined as:

(6) $$GFI = 1 - [\ (tr\ (E^{-1} \times S - I)^2 / (tr\ E^{-1} S)^2\ ],$$

(7) $$AGFI = 1 - [p \times (p+1)/2df] \times (1 - GFI).$$

Information Criterion. Akaike (1974, 1981) and Schwartz (1978) each proposed fit indices that incorporate penalty functions based on the number of parameters that are estimated. Cudeck and Browne (1983, p. 154) proposed rescaled versions of these indices expressed in terms of FF. For purposes of the present investigation, Cudeck and Browne's rescaling of the CAK (based on Akaike, 1974) and CSK (based on Schwartz, 1978) are defined as:

(8) $$CAK = FF + 2K / N,$$

(9) $$CSK = FF + (K \times \ln(N)) / N$$

where K = the number of parameters to be estimated.

The corresponding indices originally proposed by Akaike and by Schwartz are defined as:

(10) $$OCAK = X^2 + 2K,$$

(11) $$OCSK = X^2 + K \ln(N).$$

Critical N (CN). Hoelter (1983, p. 528) argued that "rather than ignoring or completely neutralizing sample size we can estimate the size that a sample must reach in order to accept the fit of a given model on a statistical basis. This estimate, referred to here as 'critical N' (CN), allows one to assess the fit of a model relative to identical hypothetical models estimated with different sample sizes." Hoelter cautioned that no firm basis could be offered as to what constituted an adequate fit, but he suggested that a value of 200 was a reasonable starting point for suggesting that differences

between the model and data may be unimportant. In practice the usefulness of CN would rest on the assumption that its value is independent of sample size. It is defines as:

(12) $CN = [[z_{crit} + (2 \times df - 1)^{1/2}]^2/[2 \times \chi^2/(N)]] + 1.$

where z_{crit} = the critical value from a normal curve table for a given probability level -- 1.96 in the present investigation.

McDonald's Fit Indices. McDonald (in press) notes that a problem with the CAK, as with many other fit indices, is that the value of the index and model selectino based on it are dependent on sample size. His DK index is based on similar formulations as the CAK but with a slightly different derivation. McDonald proposed Wald's (1943) noncentrality parameter (also see related suggestions by Steiger, 1980), rescaled to be independent of sample size, as an index of fit, estimated by:

(13) DK = FF - df/N = CAK - (2K/N) - df/N.

McDonald further proposed that DK could be transformed to yield Mc, a measure of centrality that is a consistent estimator of the asymptotic likelihood ratio scaled to be independent of sample size. Mc is scaled to lie on the interval zero to unity with unity representing a perfect fit, though sampling error may produce values greater than 1.0. It is defined as:

(14) Mc = exp (-.5 DK)

Normal Deviate Z-score. Horn and McArdle (1980) proposed the Wilson-Hilferty normal deviate Z-score (also see Bishop, Fienberg & Holland, 1975, p. 527) as a useful indicator of fit. It is defined as:

(15) $Z = [(\chi^2/df)^{1/3} - [1 - (2/9\ df)]] / [(2/9\ df)^{1/2}]$

Because this quantity is a monotonic function of χ^2 it apparently will be influenced by N so long as the hypothesized model is false.

II. Relative, Type-1 Incremental Fit Indices.

Bentler and Bonett (1980) proposed that valuable information could be obtained by comparing the ability of nested models to fit the same data. In the case of CFA it may be useful to compare the fit of the proposed target model with the fit of a null model in which all the p variables are assumed to be uncorrelated. (It should be noted that in general models for the analysis of covariance structures the null model is not the only more restrictive model that could be considered as a baseline model.) If the fit of a null model is reasonable, because the sample size is small or because the measured variables are relatively uncorrelated, then the difference in

fit between the null and target models will be small. However, if the fit of the null model is reasonable then there is little covariance to explain and no basis of support for the target model even if it also fits the data. Bentler and Bonett specifically stated that these indices are useful for comparing the fit of a particular model across samples that have unequal sizes. They cautioned that the absolute value of these indices may be difficult to interpret, but that values of less than .9 usually mean that the model can be improved substantially. Much of the value of these indices is based on the assumption that their behavior is independent of sample size.

One form of the incremental index, called type-1 incremental indices for present purposes, can be used to derive incremental fit indices from each of the stand alone indices described earlier: Absolute Value (t - n) / Maximum of (t or n), where t is the value of a stand-alone index for the target model, and n is the value for the null model. For present purposes, incremental type-1 indices were defined in relation to the FF, LHR, X^2, and X^2/df, and are denoted by appending an I1 to each stand-alone index. The X^2I2 is more commonly known as the Bentler-Bonett Index (BBI) and Bollen (1986) described an index related to the FFI1 (see Marsh, Balla & McDonald, 1988). These are defined as:

(16) $FFI1 = (FF_n - FF_t)/ (FF_N)$.

(17) $LHRI1 = (LHR_t - LHR_n)/ (LHR_t)$.

(18) $X^2I1 = BBI = (X_n^2 - X_t^2)/ (X_n^2)$.

(19) $X^2/dfI1 = (X_n^2/df_n - X_t^2/df_t)/ (X_n^2/df_n)$.

III. Parsimony Indices.

James et al. (1982) also described an alternative form of the incremental type-1 indices called the parsimony index (PI). The PI invokes a penalty function for using additional parameters by multiplying an incremental type-1 index by the ratio of the dfs for the null and target models: $PI = (df_T/df_n)$ x Incremental Type-1 Index. Using this general formulation, James et al. recommended a PI based on the X^2 defined as:

(20) $PIX^2 = (df_T/df_n) \times (X_n^2 - X_t^2)/ (X_N^2)$.

Similarly, McArdle (1986) described a parsimony index based on the RMR.

(21) $PIRMR = (df_T/df_n) \times (1 - [RMR_t / RMR_N])$.

Additional parsimony indices could be derived for other stand-alone indices,

though this might not make sense for indices that already impose a penalty function (e.g., the AGFI and the X^2/df).

IV. Relative, Incremental Type-2 Indices.

A second general form of the incremental fit indices described by Marsh, Balla, and McDonald (1988) is: Absolute Value (t - n) / Absolute Value of (e - n), where t is the value of a stand-alone index for the target model, n is the value for the null model, and e is the expected value of the stand-alone index if the target model is true. This second form of incremental index requires the expected value for a true model in addition to empirical values for the target and null models. In general, expected values for the stand-alone indices are not known for finite samples but can be estimated based on the asymptotic behavior of the indices. For example, many of the stand alone indices can be specified in terms of X^2 and the asymptotic expected value for the X^2 equals the df for the model. For purposes of the present investigation, incremental type-2 indices were derived from only the X^2/df and Dk stand-alone indices. These are denoted by appending an I2 to each of the stand-alone indices though the X^2/dfI2 is better known as the Tucker Lewis Index (Tucker & Lewis, 1973) and McDonald and Marsh (1988) refer to the DKI2 as the McDonald-Marsh Index (MMI). These are defined as:

(22) $X^2/df\text{I2} = \text{TLI} = (X_n^2/df_n - X_t^2/df_t)/(X^2_n/df_n - [1.0])$.

(23) $\text{DkI2} = \text{MMI} = (Dk_n - Dk_t) / (Dk_n - 0)$

[Note that because the expected value of Dk for a true model is 0, the incremental type-1 and type-2 forms of this index are the same.]

FIGURE CAPTIONS

FIGURE 1. Values for selected goodness-of-fit indices based on two population models (simple and complex), 8 models, and 6 sample sizes (50, 100, 200, 400, 800, and 1600 corresponding to the 6 column bars above each model respectively).

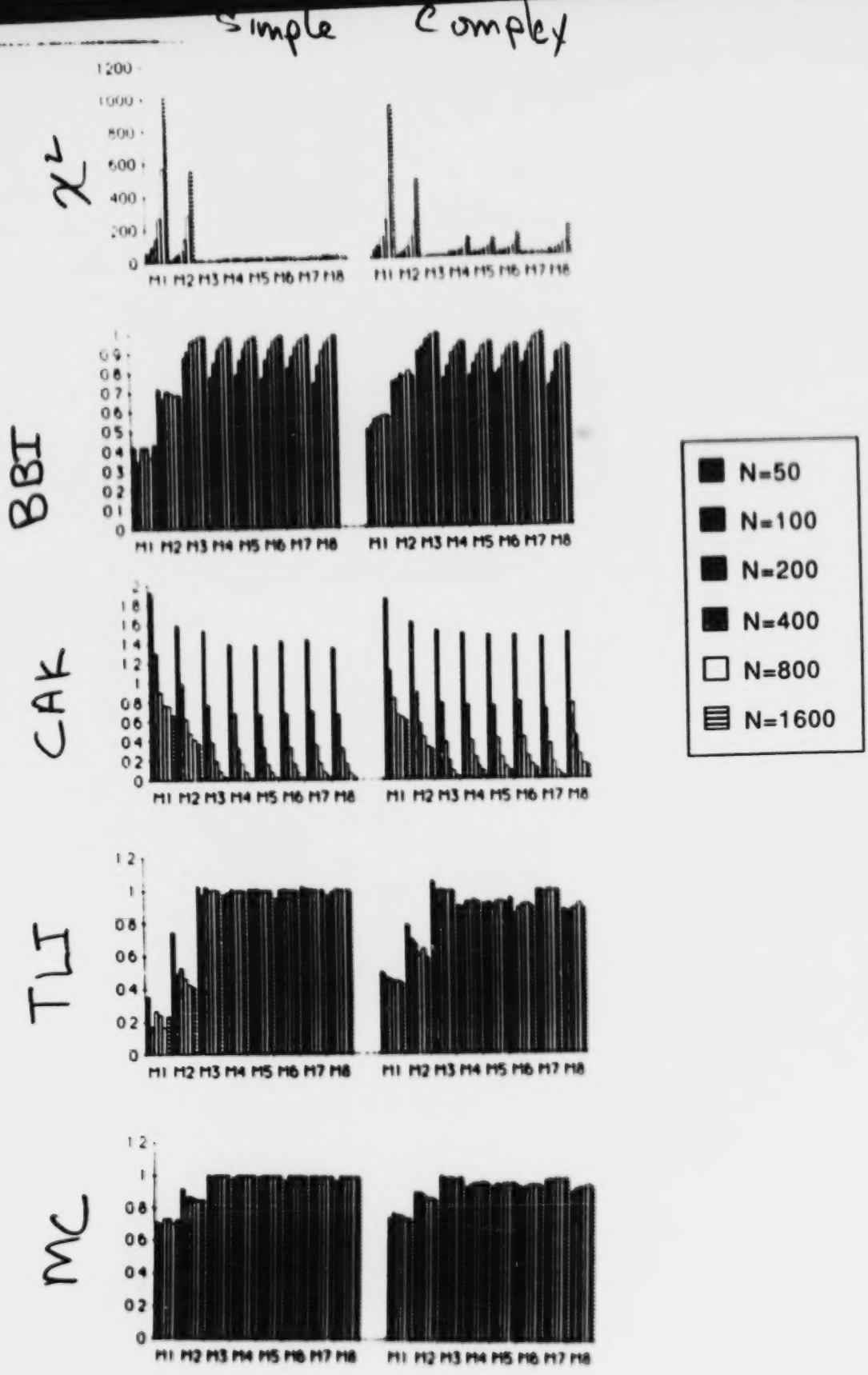
Simple
Complex
χ²
BBI
CAK
TLI
MC
N=50
N=100
N=200
N=400
N=800
N=1600
M1 M2 M3 M4 M5 M6 M7 M8

Table 1

Description of Models To Be Tested.

Abbreviation	Number of Parameters	Description
0	9	Null
1UF[a]	18	1 Unrestricted Factor
2UF[a]	26	2 Unrestricted Factors
3UF[a]	33	3 Unrestricted Factors
3F1[b]	24	3 Factors; 9 major, 3 minor factor loadings
3F2[b]	24	3 Factors; 9 major, 3 minor factor loadings
3F3[c]	24	3 Factors; 9 major, 3 minor factor loadings
3CF[d]	27	3 Complex Factors; 9 major, 6 minor factor loadings
3SF[e]	21	3 Simple Factors; 9 major factor loadings

Note. The nine models were designed to fit 9 x 9 covariance matrices generated from one of two population models. The simple simulated (SSIM) data was generated by the 3SF model in which 3 correlated factors were each defined by a unique set of three variables. Thus the most parsimonious model able to fit this data contained only 9 major factor loadings. The complex simulated (CSIM) data was generated from the 3CF model that contained three complex factors. In the 3CF model each factor was defined by three major factor loadings, the same as those in the 3SF model, and two additional minor loadings. Thus, the most parsimonious model able to fit this data contained 9 major factor loadings and 6 minor factor loadings. a -- Unrestricted factor models for 1, 2, and 3 factors. b -- For the SSIM data these models contained 3 superfluous parameters, factor loadings that had population values of zero. For the CSIM data 3 of the 6 minor factor loadings were constrained to be zero. c -- For both the SSIM and CSIM data this model contained 3 superfluous parameters. For the CSIM data all 6 minor factor loadings were constrained to be zero. d -- This model was used to generate the CSIM data. For the SSIM data it contained 6 superfluous parameters. e -- This model was used to generate the SSIM data. For the CSIM data all 6 minor (non-zero) loadings were constrained to be zero.

Table 2

Description of A Priori Contrasts To Be Tested.

	Models[a]								Predictions For[b]	
	1UF	2UF	3UF	3F1	3F2	3F3	3CF	3BF	CBIN Data	SBIN Data
Contrast 1	-1	+1	0	0	0	0	0	0	2UF > 1UF	2UF > 1UF
Contrast 2	0	-1	+1	0	0	0	0	0	3UF > 2UF	3UF > 2UF
Contrast 3	0	0	+1	0	0	0	-1	0	3UF = 3CF	3UF = 3CF
Contrast 4	0	0	+1	0	0	0	0	-1	3UF > 3BF	3UF = 3BF
Contrast 5	0	0	0	+1	+1	-2	0	0	3F1,3F2 > 3F3	3F1,3F2 = 3F3
Contrast 6	0	0	0	-1	-1	0	+2	0	3FC > 3F1,3F2	3FC = 3F1,3F2
Contrast 7	0	0	0	+1	+1	0	0	-2	3F1,3F2 > 3BF	3F1,3F2 = 3BF
Contrast 8	0	0	0	0	0	+1	0	-1	3F3 = 3BF	3F3 = 3BF
Contrast 9	0	0	0	0	0	0	+1	-1	3CF > 3BF	3CF = 3BF

Note. SBIN = simple simulated data that was generated by the 3BF model. CBIN = complex simulated data that was generated by the 3CF model.

a -- See Table 2 for a description of the models. b -- For predictions represented by > signs, models on the left side of the > signs should be better able to fit the data. For all but one prediction represented by = signs, models on the left side of the = signs have more parameters and thus provide a test of penalty functions imposed by some indices. For just contrast 5 for the SBIN data, models on both sides of the = sign have the same number of parameters.

Table 3

The Effect of Model and Sample Size on Fit Indices for SRIN Data: Effect Sizes (Etas and r's) and A Priori Contrasts

Index[a]	Model	Size[b]			Interaction	A Priori Contrasts[c]								
	Eta	Eta	r1	r2	Eta	1	2	3[d]	4[d]	5[d]	6[d]	7[d]	8[d]	9[d]
Stand-alone indices														
1 FF	.8288	.4788	-.43	-.31	.1488	+88	+88	+8	+88	+	+	+	+	+88
2 LMR	.8188	.5088	.45	.33	.1488	+88	+88	+88	+88	+	+8	+8	+	+88
3 I2	.6588	.3688	.36	.23	.3388	+88	+88	+	+88	+	+	+	+	+
4 I2/df	.6488	.3888	.38	.24	.6688	+88	+88	+	+	+	+	+	-	+
5 RMR	.8488	.4188	-.31	-.39	.2188	+88	+88	+88	+88	+	+88	+8	+8	+88
6 GFI	.8188	.5088	.45	.33	.1688	+88	+88	+88	+88	+	+8	+8	+	+88
7 AGFI	.7288	.5788	.54	.40	.1588	+88	+88	+	+	+	+	+	+	+
8 CAK	.4088	.9988	-.81	-.60	.068	+88	+88	-8	-88	+	-	-	-	-
9 CHK	.2388	.9688	-.90	-.68	.0988	+88	+88	-88	-88	+	-88	-88	-88	-88
10 BCAK	.6488	.3688	.36	.23	.6788	+88	+88	-	-88	+	-	-	-	-
11 BCHK	.5888	.5088	.48	.29	.6488	+88	+88	-88	-88	+	-88	-88	-88	-88
12 CN	.3588	.6788	.67	.40	.3988	+	+88	+	+	+	+	+	+	+
13 NK	.9388	.02	.01	.01	.10	+88	+88	+	+	+	+	+	+	+
14 NC	.9388	.02	.01	-.01	.10	+88	+88	+	+	+	+	+	+	+
15 I	.7888	.2988	.28	.20	.5288	+88	+88	+	+	+	+	+	+	+
Type-1 incremental indices														
16 FFI	.9188	.2688	.25	.20	.1888	+88	+88	+8	+88	+	+	+	+	+88
17 LMRI	.7188	.4888	.46	.28	.10	+88	+88	+	+8	+	+	+	+	+
18 I2I	.9188	.2688	.25	.19	.1888	+88	+88	+8	+88	+	+	+	+	+88
19 I2/dfI	.8788	.3488	.32	.26	.2188	+88	+88	-	+	+	+	+	-	+
Parsimony Indices														
20 PI12	.9188	.2588	.25	.19	.1988	+88	-88	-88	-88	+	-88	-88	-88	-88
21 PIRMR	.9288	.2888	.26	.20	.1488	-88	-88	-88	-88	+	-88	-88	-88	-88
Type-2 incremental indices														
22 TLI	.9288	.0788	-.05	-.04	.1488	+88	+88	+	+	+	+	+	+	+
23 NNI	.9588	.05	-.03	-.02	.1288	+88	+88	+	+	+	+	+	+	+

Note. Results are based on a series of 8 (Models) by 6 (Sample Sizes) ANOVAs conducted on the simple simulated data set. The TLI and NNI indices are based on the I^2/df and the Bt indices respectively.

8 $p < .05$; 88 $p < .01$.

a -- see Table 1 for a description of the indices. b -- Eta is the linear and nonlinear effects of sample size (N) , r1 is the linear effect of the log sample size (sample sizes are log spaced in this study), and r2 is the linear effect of sample size. c -- For each of the a priori contrasts a + sign indicates that the best fit was obtained for the model posited to fit the best, or for the model with the greatest number of parameters when the contrasted models were posited to fit equally well (see Table 3 for a description of the a priori contrasts). d -- These contrasts are between models that are equally able to fit the data.

Table 4

Effect of Sample Size and Model on Fit Indices for SSIM Data For Models 3-8

Index[a]	Model[b]		Size[c]			Interaction
	Eta	Complexity	Eta	r1	r2	Eta
Stand-alone indices						
1 FF	.18**	+	.89**	-.79	-.53	.20**
2 LHR	.18**	+	.90**	.81	.60	.19**
3 X2	.52**	+	.07	-.02	-.01	.07
4 X2/df	.05	+	.08	-.02	-.01	.10
5 RMR	.29**	+	.89**	-.86	-.70	.16**
6 GFI	.21**	+	.89**	.81	.60	.21**
7 AGFI	.05	+	.93**	.84	.63	.06
8 CAK	.05**	-	.99**	-.66	-.89	.04*
9 CSK	.11**	-	.99**	-.92	-.70	.08**
10 OCAK	.49**	-	.07	-.02	-.01	.08
11 OCSK	.49**	-	.84**	.84	.77	.13**
12 CN	.06	+	.82**	.74	.82	.07
13 DK	.07	+	.12	-.07	-.04	.14
14 MC	.07	+	.10	.05	.03	.14
15 Z	.05	+	.09	-.03	-.02	.09
Type-1 incremental indices						
16 FFI1	.21**	+	.84**	.80	.63	.19**
17 LHRI1	.09	+	.66**	-.56	-.39	.09
18 X2I1	.21**	+	.84**	.80	.63	.19**
19 X2/dfI1	.03	+	.86**	.83	.65	.05
Parsimony Indices						
20 PIX2	.89**	-	.38**	.37	.29	.14**
21 PIRMR	.95**	-	.29**	.28	.21	.10**
Type-2 incremental indices						
22 TLI	.06	+	.06	.03	.01	.03
23 MMI	.06	+	.02	.01	.00	.06

Note. Results are based on a series of 6 (Models) by 6 (Sample Sizes) ANOVAs conducted on the simple simulated data set. For purposes of these analyses only the three-factor models, all of which are able to fit the data, were included.

* $p < .05$; ** $p < .01$.

a -- see Table 1 for a description of the indices. b -- Because all models are equally able to fit the data, differences between models are a test of the relations between each model and model complexity. Under the Complexity column a + indicates that fit improved with the addition of superfluous parameters and a - indicates that fit was poorer with the addition of superfluous parameters. c -- Eta is the linear and nonlinear effects of sample size (N) , r1 is the linear effect of the log sample size (sample sizes are log spaced in this study), and r2 is the linear effect of sample size.

Table 9

The Effect of Model and Sample Size on Fit Indices for CBIH Data: Effect Sizes (Eta and r2) and A Priori Contrasts

Index[a]	Model	Size[b]			Interaction	A Priori Contrasts[c]								
								[d]					[d]	
	Eta	Eta	r1	r2	Eta	1	2	3	4	5	6	7	8	9
Stand-alone indices														
1 FF	.73##	.57##	-.51	-.36	.16##	+##	+##	+#	+##	+	+##	+##	+##	+##
2 LMR	.74##	.58##	.52	.38	.14##	+##	+##	+##	+##	+	+##	+##	+##	+##
3 I2	.60##	.49##	.44	.49	.62##	+##	+##	+	+##	+##	+##	+##	+##	+##
4 I2/df	.57##	.53##	.47	.53	.61##	+##	+##	+	+##	+##	+##	+##	+	+##
5 RMR	.80##	.46##	-.43	-.32	.19##	+##	+##	+##	+##	+	+##	+##	+##	+##
6 GFI	.74##	.57##	.50	.36	.16##	+##	+##	+##	+##	+	+##	+##	+##	+##
7 AGFI	.61##	.68##	.61	.44	.15##	+##	+##	+	+##	+	+##	+#	+	+##
8 CAK	.30##	.94##	-.84	-.62	.05	+##	+##	-	+#	+	+##	+	+	+##
9 CBK	.16##	.98##	-.91	-.69	.09##	+##	+##	-##	-##	+	-	-	-#	-#
10 OCAK	.57##	.50##	.45	.50	.63##	+##	+##	-	-##	+##	+##	+##	+	+##
11 OCBK	.49##	.64##	.60	.63	.58##	+##	+##	-##	-	+##	+##	+	-	-##
12 CN	.53##	.48##	.46	.48	.61##	+	+##	+	+##	+	+##	+	+	+##
13 NK	.08##	.06	-.02	-.01	.11	+##	+##	+	+##	+	+##	+##	+	+##
14 NC	.08##	.05	.02	.00	.12	+##	+##	+	+##	+	+##	+##	+	+##
15 I	.67##	.56##	.54	.55	.45##	+##	+##	+	+##	+##	+##	+##	+##	+##
Type-1 incremental indices														
16 FFI	.07##	.37##	.35	.27	.15##	+##	+##	+##	+##	+	+##	+##	+##	+##
17 LMRI	.60##	.51##	.35	.22	.09	+##	+##	+	+##	+	+#	+	+	+##
18 I2I	.07##	.37##	.35	.27	.15##	+##	+##	+##	+##	+	+##	+##	+##	+##
19 I2/dfI	.79##	.47##	.45	.35	.18##	+##	+##	+	+##	+	+##	+##	+	+##
Parsimony indices														
20 PI2	.07##	.34##	.33	.25	.16##	-	-##	-##	-##	+	-##	-##	-##	-##
21 PIRMR	.91##	.32##	.28	.20	.14##	-##	-##	-##	-##	-	-##	-##	-##	-##
Type-2 incremental indices														
22 TLI	.85##	.07	-.04	-.03	.12	+##	+##	+	+##	+	+##	+#	+	+##
23 RNI	.90##	.04	-.03	-.02	.10	+##	+##	+	+##	+	+##	+#	+	+##

Note. Results are based on a series of 8 (Models) by 6 (Sample Sizes) ANOVAs conducted on the complicated simulated data set.

a -- see Table 1 for a description of the indices. b -- Eta is the linear and nonlinear effects of sample size (N) , r1 is the linear effect of the log sample size (sample sizes are log spaced in this study), and r2 is the linear effect of sample size. c -- For each of the a priori contrasts a + sign indicates that the best fit was obtained for the model posited to fit the best, or for the model with the greatest number of parameters when the contrasted models were posited to fit equally well (see Table 3 for a description of the a priori contrasts). d -- These contrasts are between models that are equally able to fit the data.

# p < .05; ## p < .01.